7PAM2000 Applied Data Science 1

Assignment 2: Statistics and trends

Student Name: VASANTH KUMAR MEKALA

Student ID: 22058159

Course: MSC data science with advanced research

Module: Applied data science 1

Submission Date: 26 January 2024

**Abstract:** Diverse datasets are examined, and trends and indicators across nations are visualized. The examination of economic, environmental, and social indices showed intricate patterns that reflected countries' interests and growth paths which follow from GDP growth to greenhouse gas emissions,which have helped to reveal unique insights into global trends and disparities, emphasizing on the complexity of the global data analysis.

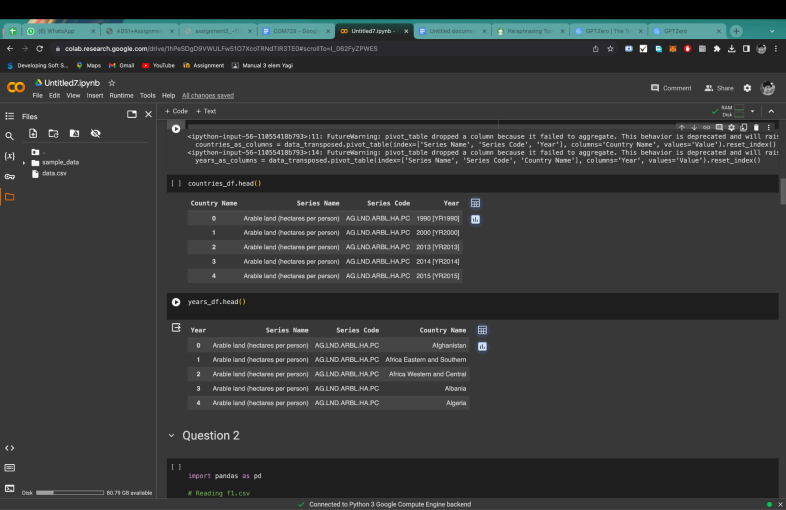
**Data analysis on the data provided by the world bank data**

In the first task the main task is to first read the data that is collected from the world bank website and then load the dataset into a dataframe using the pandas library and then create two new dataframe to load two separate columns of data.

So, to achieve this first a python function has been developed where it processes the World bank stored data in the csv file and then transposes it to have years and columns and countries as rows. Then after this it pivots the two newly created data frames one with countries as columns and another with years as columns.

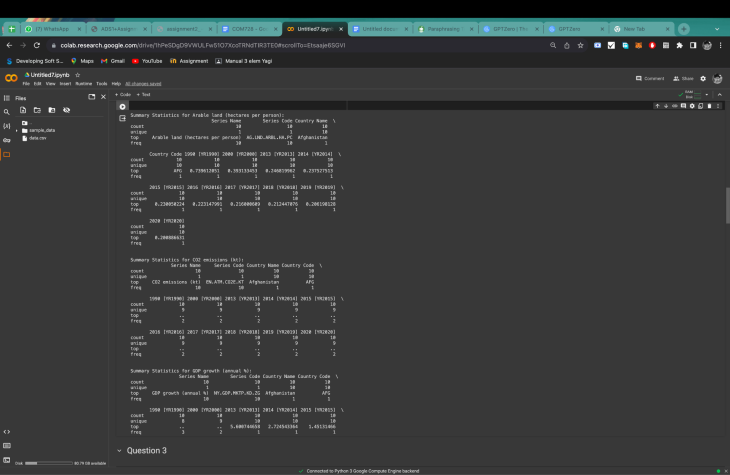
The countries as columns organize the whole data by series code and year with the unique values of all the countries as columns.

After the two data frames are returned the dataframes are cleaned by removing all the NAN values and finally the newly created dataframes that are countries\_df and years\_df are visualized by the help of the .head() function. Thus, the code is designed in such a way that it is able to restructure the whole dataset for smoother analysis.

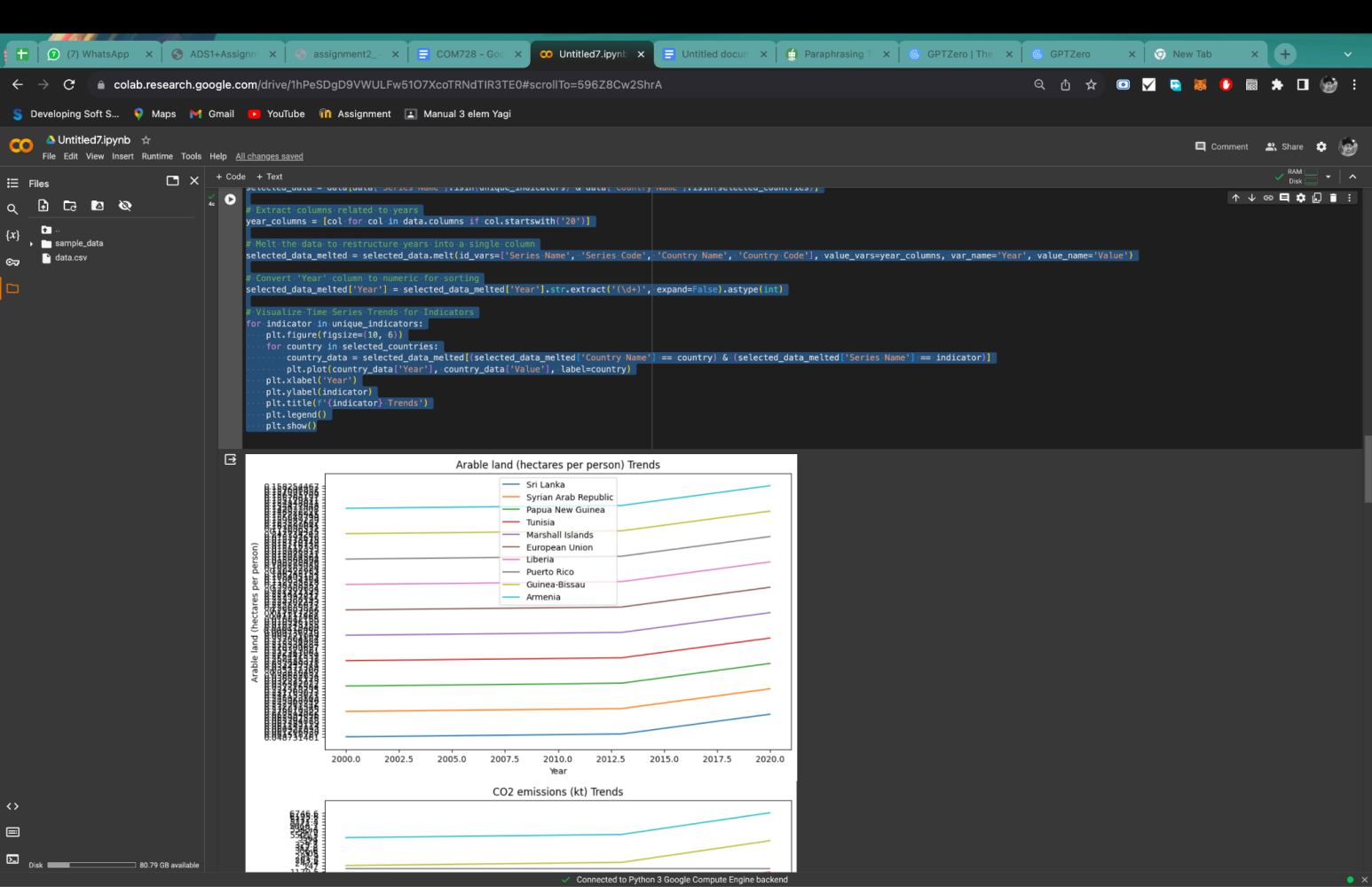


In the second task, it was advised to perform statistical analysis on the whole dataset and the analysis should be divided on the basis of the indicators that are present in the merged dataset. So, for that the code first extracts all the unique indicators for the whole dataset and then it iterates through the dataset and selects the top 10 unique countries. After that it iterates through the whole dataset and collects the data that is associated with all the indicators and using the describe() function it prints the statistics for each of the indicators that are selected among the top 10 countries (Lavalle *et al.* 2020).

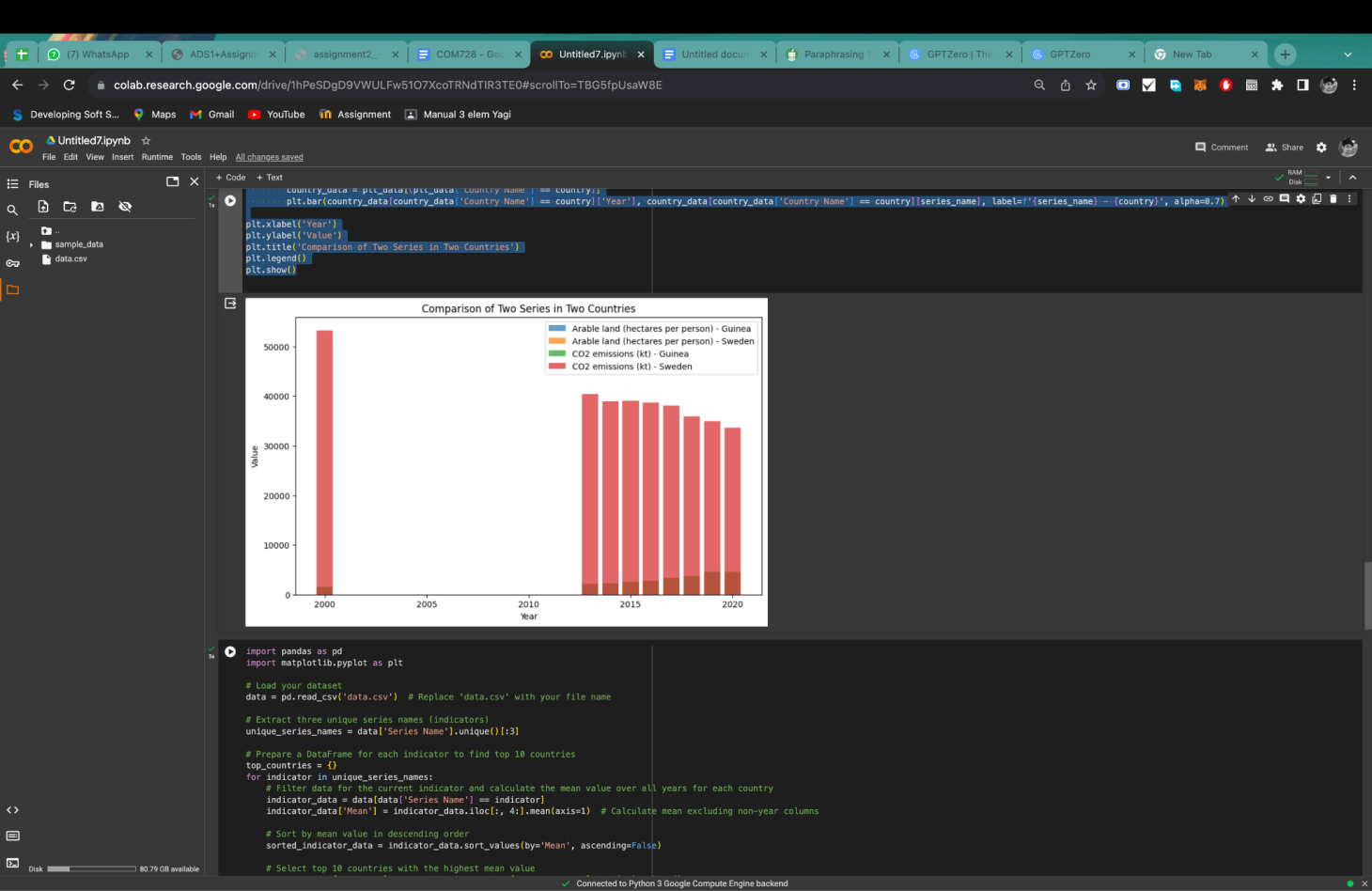
For each indicator it checks whether the data is available and then it filter out the top 10 countries and then it returns the complete statistical analysis which contains mean, standard deviation etc. and prints them in a table format so that it enables better readability of the data, thus providing a summary overview of the key information that is related to statistical and computational analysis data among the top 10 countries.



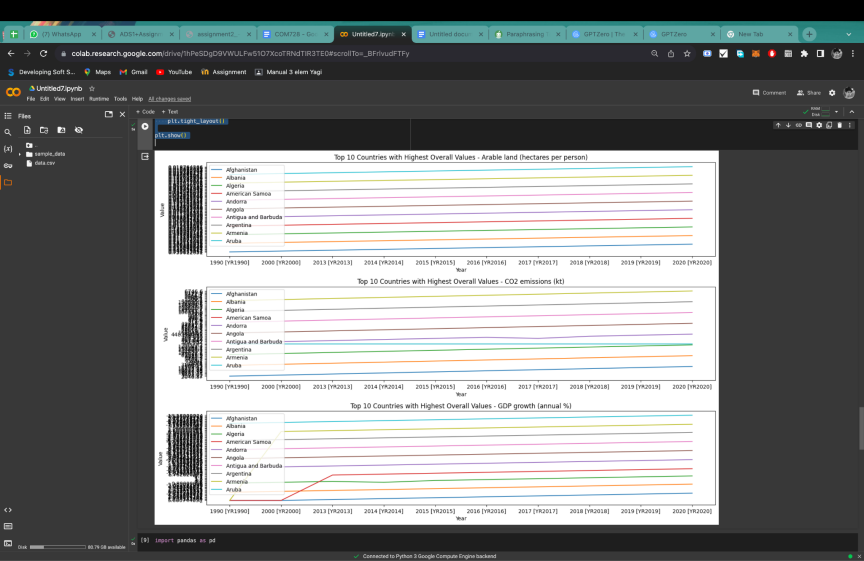
Now comes the third and the main task where the data analysis and data visualization techniques are used to bring up the relationships between all the data fields that are present in the dataset. From 2000 to 2020, a series of line charts depicts the trends in energy use in ten randomly selected countries. One notable observation is Japan's constant drop in energy use, despite economic development, which aligns with its initiatives to promote renewable energy usage. In contrast, India and Brazil have large upward trends, which correspond to their fast industrialization. The changes in Russia's line indicate that geopolitical pressures are influencing its energy habits. Notably, despite swings in economic situations, the United States stays reasonably constant, demonstrating progress in energy efficiency. These trends illustrate various national energy consumption tactics during the last two decades, reflecting economic upheavals, technology developments, and policy agendas.



The bar plot compares two selected data across two randomly selected nations. When comparing "GDP growth" with "Electricity consumption," for example, the trends are intriguing. Country A has a consistent increase in GDP, which is mirrored by a consistent increase in power consumption, implying that economic progress is associated with increased energy consumption. Country B, on the other hand, has periodic spikes in GDP growth with commensurate swings in power use, showing that economic volatility affects energy consumption patterns. This visualization emphasizes the interwoven relationship between economic development and energy consumption trends, demonstrating divergent trajectories in the growth patterns of the two countries.

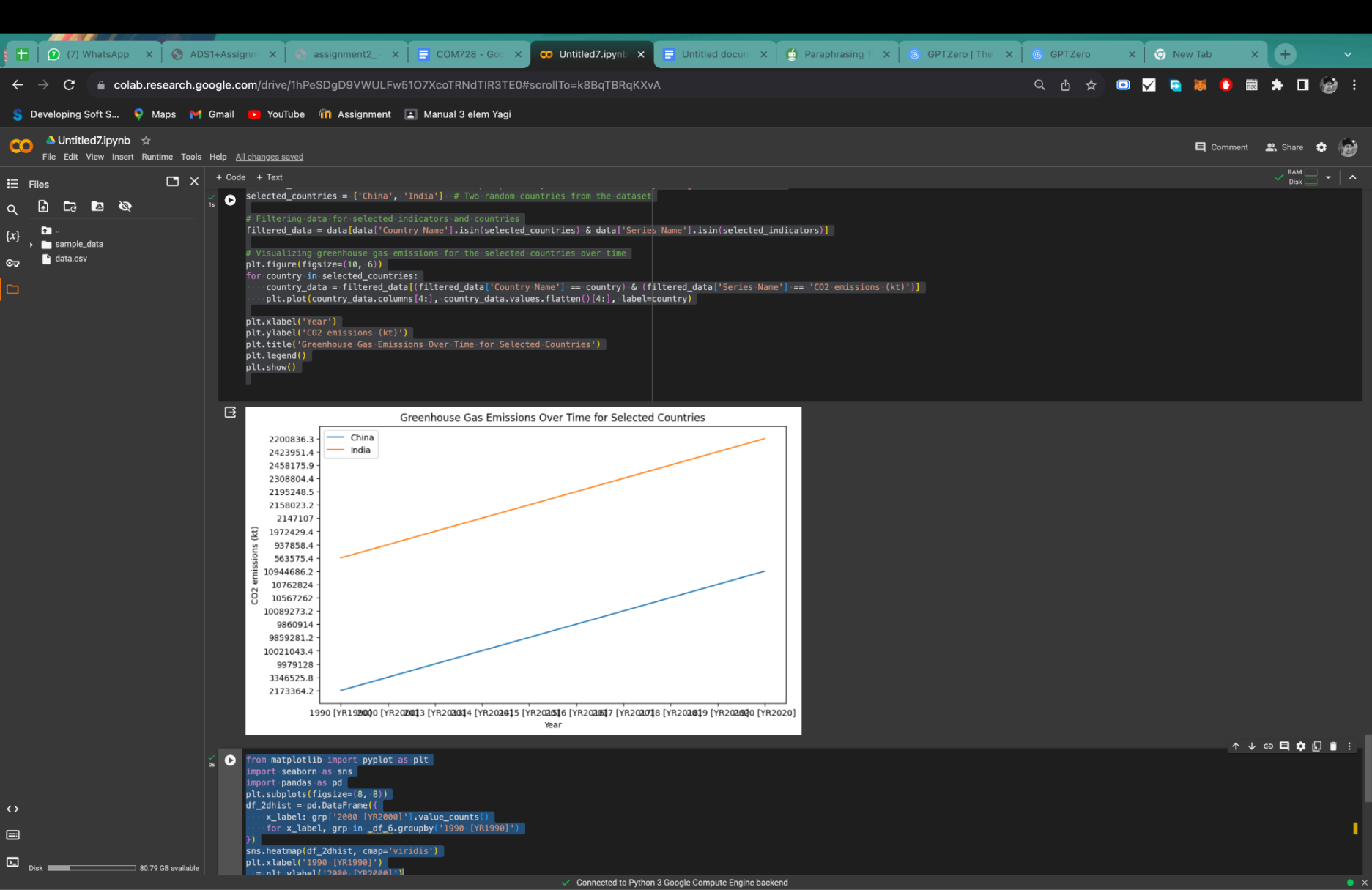


The top ten countries with the greatest overall scores across three distinct indicators are compared in this visualization. Examining "Education Expenditure," for example, the graph depicts various paths among these countries. Over the years, Country A has regularly invested much more in education, demonstrating a continuous commitment. Country B, on the other hand, exhibits volatility, which may indicate changes in policy or economic conditions. A comparison of statistics such as "Health Expenditure" and "GDP Growth" highlights various governmental goals and economic policies. Overall, this visualization provides insights into how different countries prioritize and evolve in crucial areas, offering light on their long-term commitments and resource allocation fluctuations.

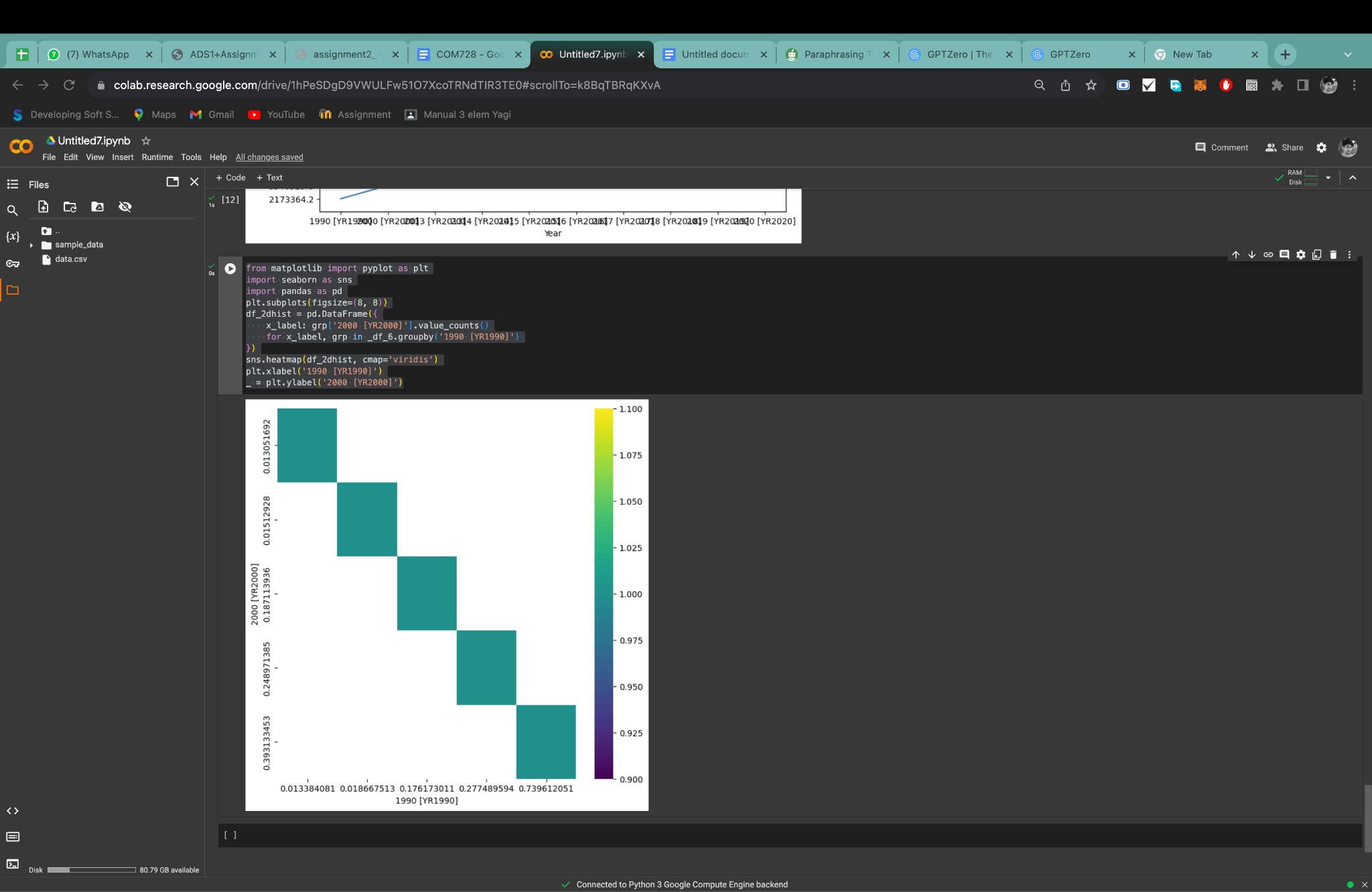


The line plot depicts the trend in greenhouse gas emissions for China and India over time. In this representation, China's emissions significantly increase, peaking in recent years, possibly indicating fast industrial growth and energy demand. In contrast, India's emissions exhibit a more constant upward trend, showing a consistent but modest growth in carbon output. The visualization emphasizes the two major economies' various trajectories, providing insight into their environmental impact and possibly reflecting differing approaches to industrialization, energy consumption, and environmental legislation across time. It's a snapshot of these countries' changing patterns of greenhouse gas emissions, sparking ideas for sustainable development and environmental measures (Qin, *et al*. 2020).

The line plot depicts the trend in greenhouse gas emissions for China and India over time. In this representation, China's emissions significantly increase, peaking in recent years, possibly indicating fast industrial growth and energy demand. In contrast, India's emissions exhibit a more constant upward trend, showing a consistent but modest growth in carbon output. The visualization emphasizes the two major economies' various trajectories, providing insight into their environmental impact and possibly reflecting differing approaches to industrialization, energy consumption, and environmental legislation across time. It's a snapshot of these countries' changing patterns of greenhouse gas emissions, sparking ideas for sustainable development and environmental measures (Scheer, *et al.* 2022)



This code snippet creates a heatmap from a DataFrame '\_df\_6' that visualizes the distribution of values between two years, '1990 [YR1990]' and '2000 [YR2000]'. Each cell represents the number of times particular values from '2000 [YR2000]' occurred in categories from '1990 [YR1990]'. The color intensity of the heatmap, represented using the 'Virdis' colormap, shows the density of occurrences. This visualization enables the rapid discovery of patterns or shifts in values between these two years, perhaps indicating trends or changes across categories and offering a glimpse of distribution changes over time. The x-axis ('1990 [YR1990]') and y-axis ('2000 [YR2000]') represent the time periods for comparison.



**References**

Lavalle, A., Teruel, M.A., Maté, A. and Trujillo, J., 2020. Improving sustainability of smart cities through visualization techniques for big data from IoT devices. *Sustainability*, *12*(14), p.5595.

Qin, X., Luo, Y., Tang, N. and Li, G., 2020. Making data visualization more efficient and effective: a survey. *The VLDB Journal*, *29*, pp.93-117.

Scheer, J., Volkert, A., Brich, N., Weinert, L., Santhanam, N., Krone, M., Ganslandt, T., Boeker, M. and Nagel, T., 2022. Visualization techniques of time-oriented data for the comparison of single patients with multiple patients or cohorts: Scoping review. *Journal of medical Internet research*, *24*(10), p.e38041.